AD_________________

Award Number: W81XWH-13-1-0335

TITLE: Whole Genome Sequencing of High-Risk Families to Identify New Mutational Mechanisms of Breast Cancer Predisposition

INITIATING PRINCIPAL INVESTIGATOR: Tom Walsh, PhD

CONTRACTING ORGANIZATION: University of Washington
Seattle, WA, 98195

REPORT DATE: December 2015

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

□ Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* December 2015	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 15 Sep 2013 - 14 Sep 2015

4. TITLE AND SUBTITLE Whole Genome Sequencing of High-Risk Families to Identify New Mutational Mechanisms of Breast Cancer Predisposition	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER W81XWH-13-1-0335
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Tom Walsh, PhD, Mary-Claire King, PhD	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Washington Seattle, WA, 98195	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

As genes for inherited disease are increasingly well characterized by next generation sequencing approaches, it is clear that some mutations may act through promoters, enhancers, and other non-coding regulatory regions. Our hypothesis for this proposal is that much of the substantial remaining familial risk of breast cancer is due to a large number of individually rare alleles of moderate-to-severe effect located in the non-coding regions of the genome. For this proposal we will evaluate 30 large, extended kindreds severely affected with breast cancer, each of whom has been comprehensively evaluated in our lab by targeted genomic sequencing for mutations of all classes in all known breast cancer genes and by whole exome sequencing for coding region mutations exome-wide. These families are a unique discovery series for identification of regulatory mutations that may reveal new mutational mechanisms for breast cancer predisposition.

15. SUBJECT TERMS

Genome sequencing, mutation, breast cancer susceptibility genes

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	139 words	10	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

1. INTRODUCTION: Despite tremendous advances in mutation detection with gene panels and exome sequencing, most families severely affected by breast cancer do not have causative alleles identified from the protein-coding region of the genome. We hypothesize that the causal mutations in many of these families lie in uncharacterized regulatory regions of the genome. Through whole genome sequence analysis of severely affected families and functional annotation and experimental evidence, we undertook to identify new mutational mechanisms that predispose to breast cancer. Our ultimate goal is to enable information on newly identified mutations and mutational mechanisms to be useful to clinicians and to women and their families.

2. KEYWORDS: Breast cancer, *BRCA1*, *BRCA2*, whole genome sequencing, promoter, enhancer, transcription factor binding site, gene regulation, mutation.

3. ACCOMPLISHMENTS:

What were the major goals of the project?

The major goals of this project as stated in the approved SOW are listed below:

TASK 1. (Walsh) Perform whole genome sequencing of germline DNA from 100 breast cancer patients selected from 30 severely affected families.

1a. Prepare 100 standard paired end library with 300-400bp inserts (months 1-3)
1b. Prepare 100 mate-paired library with tightly defined 6kb inserts (months 1-3)
1c. Sequence the paired end and mate-paired libraries on a HiSeq2500 (months 2-9)

All components of TASK 1 have been completed

TASK 2. (Walsh) Annotating sequencing genome variants with respect to population frequency and overlap with ENCODE regions.

2a. Align reads to the reference sequence (months 4-10)
2b. Identify SNPs, indels, CNVs and rearrangements by bioinformatics tools (months 4-10)
2c. Filter variants from Task 2b against public databases to remove common events (months 4-10)
2d. Filter rare and private variants from Task 2c within families to obtain segregating variants (months 4-10)
2e. Compare surviving events from Task 2d to ENCODE regions (months 4-10)
2f. Filter variants from Task 2e to ENCODE variants mapped only in breast tissues/lines (months 4-10)

All components of TASK 2 have been completed

TASK 3. (King) Characterize potential regulatory variants.

3a. Generate enhancer constructs with wildtype and variant regulatory regions (months 8-16)
3b. Transfect constructs into cell lines, monitor luciferase activity (months 8-16)
3c. Measure gene expression in patients' lymphoblasts (months 8-16)

All components of TASK 3 have been completed

TASK 4. (King) Resequence mutant regulatory regions in large series of patients to identify additional mutations

4a. Design molecular inversion probes (MIPs) for promising regulatory regions (months 12-24)
4b. Perform MIP amplification, hybridization and sequencing (months 12-24)
4c. Annotate variants within regulatory regions with respect to frequency (months 12-24)
4d. Statistical analysis of variants (months 12-24)

All components of TASK 4 have been completed

What was accomplished under these goals?

TASK 1: (Walsh) We prepared both library types and generated whole genome sequencing data on the 100 breast cancer patients. We obtained median sequence coverage of 34X, with 99.89% of basepairs read at least 8 times. Figure 1 below illustrates the distribution of read depths across the 100 samples and Figure 2 shows the distribution of median read depths.

Figure 1. Distribution of read depths across the 100 whole genome sequenced samples.

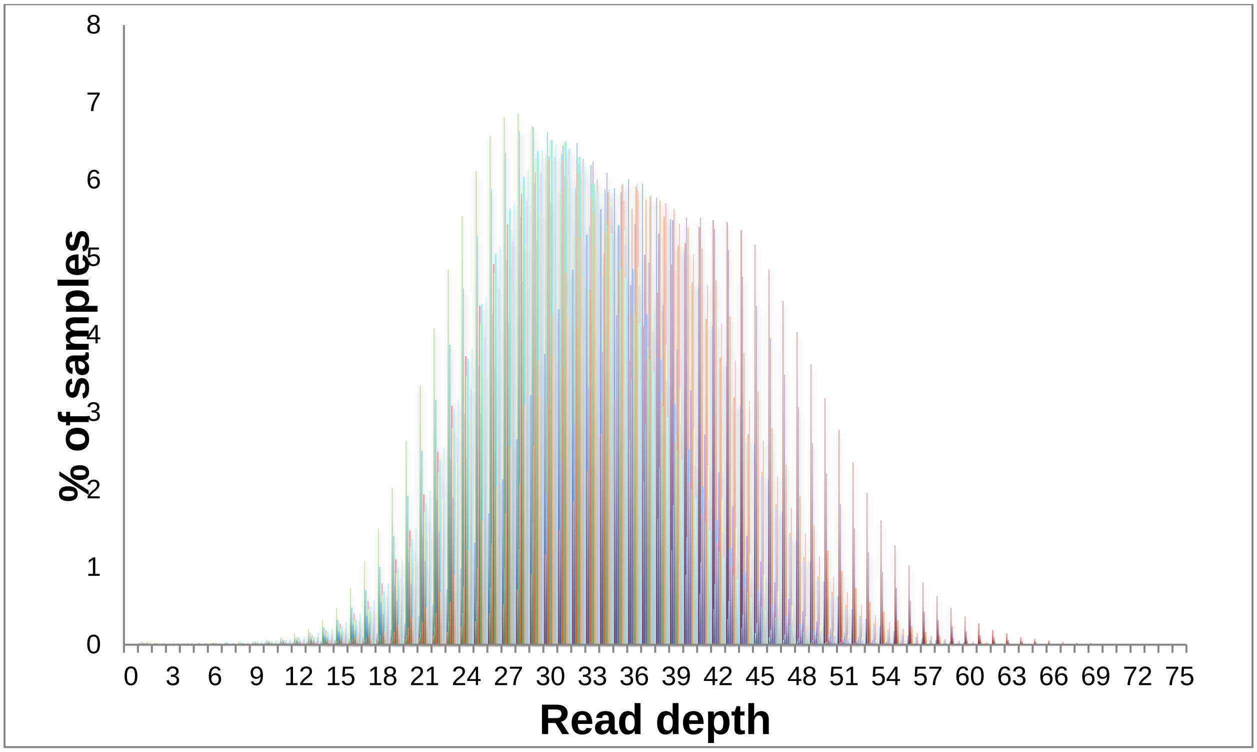
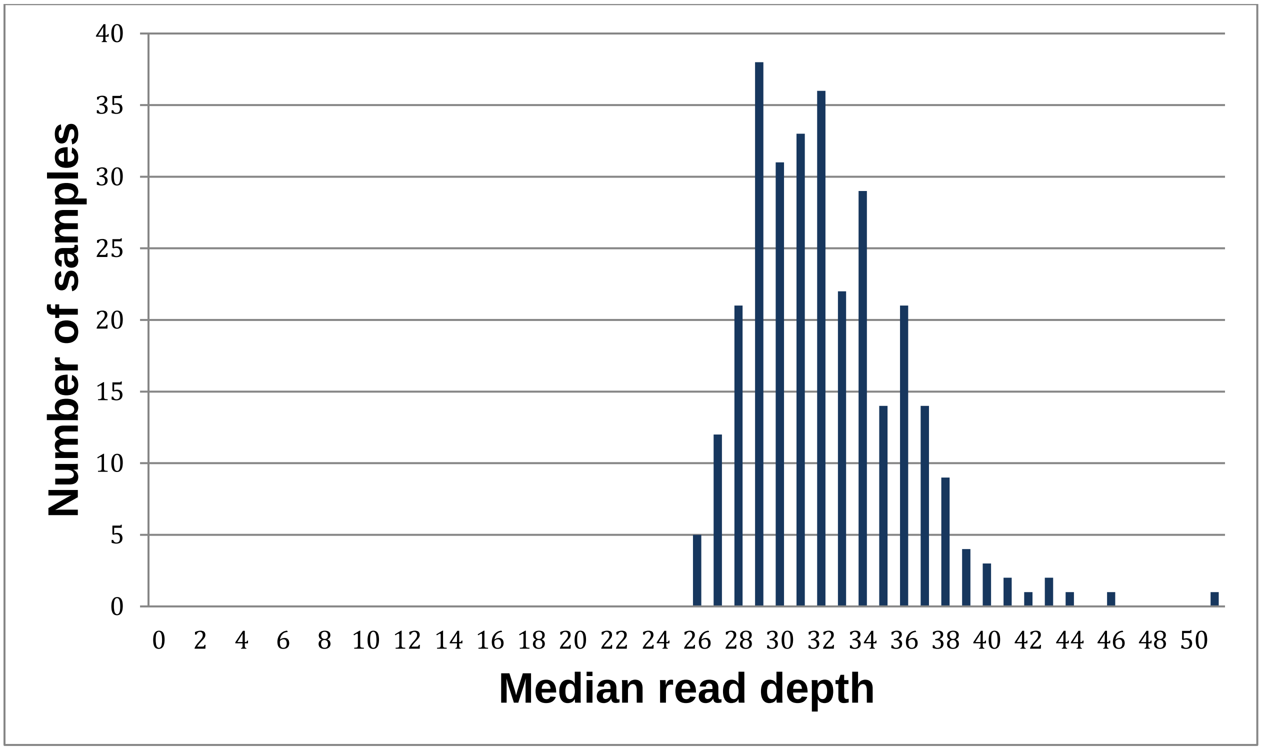


Figure 2. Distribution of median read depths across the 100 whole genome sequenced samples.



TASK 2: (Walsh) In Year 2 after sequencing was completed, we developed a functional annotation approach to filter variants from the whole genome sequences. Sequence reads from TASK1 were aligned to the genome with BWA[1] which is presently the preferred alignment tool of the genomics community. Single nucleotide variants (SNVs) and small indels (1-50bp) we identified by GATK[2], UnifiedGenotyper[3], FreeBayes[4] and Samtools[5]. Structural Variants (SVs) greater than 500bp were identified with Lumpy[6], SVtyper[7] and CNVnator[8]. We felt it was important to use all these calling tools since they have different specificities and sensitivities.

In order to remove common and likely benign SNPs, indels and SVs for follow up in TASK3 we filtered using these criteria:

(i) Less than 1% frequency in the 1000 Genomes Project[9] and Genomes of the Netherlands[10]
(ii) Present in less than 4 of the breast cancer families that we sequenced in TASK1
(iii) Variant read proportion greater than 0.25
(iv) Read depth greater than 8 in at least one sample

The Table below shows variant data from Family 1041 categorized by functional effect.

Table 1. Distribution of different types of mutations at different filtering levels in the whole genome sequencing data of two patients from a severely affected breast cancer Family 1041.

	All	Shared	Rare	Excluding IBD0
Intergenic	3,345,727	1,650,045	35,927	3,990
ncRNA	266,300	130,836	3,104	329
Up-downstream	72,754	35,055	865	98
Untranslated	48,884	23,492	435	49
Intronic	1,986,665	970,436	20,088	2,139
Missense	13,933	6,891	48	11
Silent	15,150	7,567	30	5
Nonsense	116	45	1	0
Splice	198	120	1	0
Stoploss	15	8	0	0
In-frame	385	162	2	0
Frame-shift	199	92	2	1
Total	5,750,326	2,824,749	60,503	6,622

In order to prioritize SNPs, indels and SVs for follow up in TASK3 we filtered remaining variants by segregation in each family. We excluded any variant that was not present in at least three women with breast cancer in a family.

The remaining variants were annotated for regulatory potential as follows:

1. ENCODE regions[11]: We used chromatin predictions from Human Mammary Epithelial cell lines (DNaseI and H3K27Ac signals). In addition, transcription factor CHiP-seq and DNaseI clusters from 7 breast cancer cell lines and finally predicted Transcription Factor binding motifs.

2. Potential motif disruption: For each potential ENCODE region and overlapping Transcription Factor binding motifs, we calculated motif score changes using position weight matrices

3. Non coding prediction: We used scores generated by DANN[12], FATHMM[13] and FunSeq2[14] that combine population frequencies and nucleotide conservation for non-coding regions of the genome.

At the end of TASK 2 we selected 112 non-coding variants from the 33 breast cancer families that shared these features:

- (i) segregated with breast cancer in a family
- (ii) private or ultra-rare in population databases
- (iii) predicted to disrupt a potential regulatory element
- (iv) within 2MB of a known breast cancer predisposition gene

TASKS 3 and 4 were performed in Year 2.

TASK 3. (King) We assessed gene expression levels by RT-PCR and targeted RNAseq using RNA from freshly acquired blood samples of the 33 patients harboring the 112 variants described in TASK 2. Of the 112 variants evaluated 11 showed expression profiles suggesting dysregulation of gene transcription.

TASK 4. (King) We genotyped the 11 variants from TASK3 in 960 female cancer free controls and 960 unrelated breast cancer patients different from those used in TASK1. We did not observe these variants in any of the 960 controls or in any other familial breast cancer cases confirming they are rare or potentially private to their host families.

We also sequenced approximately 300bp around each of the 11 variants of interest in 960 female cancer-free controls and 960 unrelated breast cancer patients. There were no additional rare variants that were predicted to disrupt a regulatory element.

What opportunities for training and professional development has the project provided?

Nothing to report

How were the results disseminated to communities of interest?

We are in the process of writing two manuscripts describing:

- (i) Our approach for targeted RNA sequencing to evaluate non-coding variants
- (ii) Our bioinformatics analysis of family based whole genome sequencing data

Our main findings describing the 11 interesting non-coding variants will require additional confirmation before publishing and presentation at scientific meetings.

What do you plan to do during the next reporting period to accomplish the goals?

Nothing to report (end of this project)

4. IMPACT

What was the impact on the development of the principal discipline(s) of the project?

Our approach integrated whole genome sequencing with experimental biology and with application and development of bioinformatics tools to discover regulatory variants that may predispose women to develop breast cancer. Through our publications and data we hope that this approach will encourage other researchers to examine non-coding portions of the genome for additional cancer predisposing mutations.

What was the impact in other disciplines? Nothing to report

What was the impact on technology transfer? Nothing to report

What was the impact on society beyond the science and technology?

Through our publications and data we hope that non-coding variants that definitely impact breast cancer gene expression will be incorporated into clinical care.

5. CHANGES/PROBLEMS

Changes in approach and reasons for change Nothing to report

Actual or anticipated problems or delays and actions or plans to resolve them Nothing to report

Changes that had a significant impact on expenditures Nothing to report

Significant changes in the use or care of human subjects, vertebrate animals, biohazards, and/or select agents Nothing to report

6. PRODUCTS

Publications, conference papers, and presentations

Nothing to report at present, although we have two manuscripts in preparation

Websites or other Internet sites Nothing to report

Technologies or techniques Nothing to report

Inventions, patent applications, and/or licenses Nothing to report

Other products Nothing to report

7. PARTICIPANTS AND OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name:	Tom Walsh, PhD
Project Role	PI
Research Identifier	
Nearest person month worked	24
Contribution to Project	Tom Walsh performed the whole genome sequencing and participated in variant filtering and selection
Funding Support	NIH

Name:	Mary-Claire King, PhD
Project Role	PI
Research Identifier	
Nearest person month worked	24
Contribution to Project	Mary-Claire King performed candidate variant selection and supervised the expression analysis experiments
Funding Support	ACS, NIH, Breast Cancer Research Foundation, Komen for the Cure

Name:	Ming Lee, PhD
Project Role	Bioinformaticist
Research Identifier	
Nearest person month worked	24
Contribution to Project	Ming Lee performed the whole genome sequencing data analysis and downstream bioinformatics and variant analysis
Funding Support	NIH

Name:	Silvia Casadei, PhD

Project Role	Research Scientist
Research Identifier	
Nearest person month worked	24
Contribution to Project	Silvia Casadei performed the gene expression analysis, variant genotyping and candidate region sequencing
Funding Support	NIH

Has there been a change in the active other support of the PD/PIs or senior/key personnel since the lat reporting period?

No change

What other organizations were involved as partners?

None

8. SPECIAL REPORTING REQUIREMENTS

COLLABORTIVE AWARDS

A duplicate report with the Tasks clearly marked with the responsible PI was submitted for this award.

APPENDICES

References cited:

[1] Li H, Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinform 25, 1754-1760.

[2] The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. (2010). McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, Genome Res *20:*1297-1303

[3] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491-498.

[4] Garrison E, Marth G (2012). Haplotype-based variant detection from short-read sequencing. http://arxiv.org/abs/1207.3907

[5] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25: 2078-2079.

[6] Layer RM, Chiang C, Quinlan AR, Hall IM. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15:R84.

[7] Hall I. Bayesian genotyper for structural variants. https://github.com/hall-lab/svtyper

[8] Abyzov A, Urban AE, Snyder M, Gerstein M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21:974-984.

[9] 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65.

[10] Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 46:818-825.

[11] Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC. Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA. 111:6131-6138.

[12] Quang D, Chen Y, Xie X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 31:761-763.

[13] Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, Campbell C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 31:1536-1543.

[14] Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. 15:480.